

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-306594

(43)Date of publication of application : 02.11.2001

(51)Int.Cl.

G06F 17/30

(21)Application number : 2000-117405

(71)Applicant : MITSUBISHI ELECTRIC CORP

(22)Date of filing : 19.04.2000

(72)Inventor : TSUDAKA SHINICHIRO

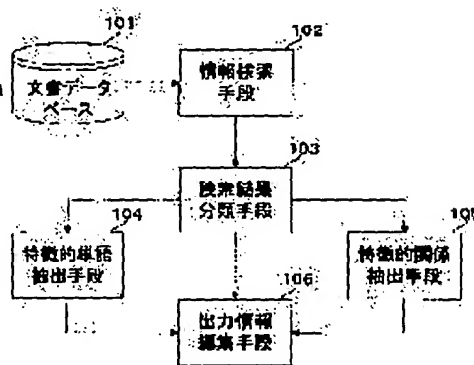
ARITA HIDEKAZU

KONAKA HIROYOSHI

(54) INFORMATION RETRIEVAL DEVICE AND STORAGE MEDIUM STORED WITH INFORMATION RETRIEVAL PROGRAM**(57)Abstract:**

PROBLEM TO BE SOLVED: To provide an information retrieval device for solving the problem in the conventional device that a user cannot acquire knowledge efficiently with the result of information retrieval as a starting point since the result of ordinary information retrieval is presented merely as a list of documents.

SOLUTION: While defining a document database 101 as a retrieval target, an information retrieving means 102 performs basic information retrieval, on the basis of the information request of the user, document groups on the acquired list of the retrieved result are sorted into document sets composed of mutually similar documents by a retrieved result sort means 103, a characteristic word and a characteristic relation are extracted from each of document sets of the retrieved result by a characteristic word extract means 104 and a characteristic relation extracting means 105 and on the basis of the sorted groups, the extracted word and relation, the information of an output picture operable for the user is generated by an output information-editing means 106.

**LEGAL STATUS**

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2001-306594

(P2001-306594A)

(43) 公開日 平成13年11月2日(2001.11.2)

(51) Int.Cl. ⁷	識別記号	F I	テームコード(参考)
G 0 6 F 17/30	2 1 0	G 0 6 F 17/30	2 1 0 D 5 B 0 7 5
	1 7 0		2 1 0 A
	3 5 0		1 7 0 A
			3 5 0 C

審査請求 未請求 請求項の数12 O L (全 11 頁)

(21) 出願番号 特願2000-117405(P2000-117405)

(22) 出願日 平成12年4月19日(2000.4.19)

(71) 出願人 000006013

三菱電機株式会社

東京都千代田区丸の内二丁目2番3号

(72) 発明者 津高 新一郎

東京都千代田区丸の内二丁目2番3号 三
菱電機株式会社内

(72) 発明者 有田 英一

東京都千代田区丸の内二丁目2番3号 三
菱電機株式会社内

(74) 代理人 100093562

弁理士 児玉 俊英

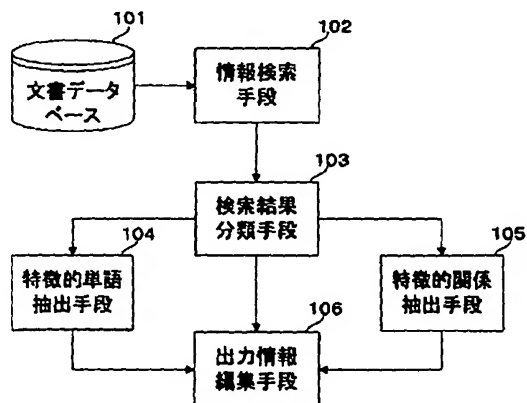
最終頁に続く

(54) 【発明の名称】 情報検索装置及び情報検索プログラムを格納した記憶媒体

(57) 【要約】

【課題】 通常の情報検索の結果は単なる文書のリストによってしか提示されないため、ユーザは情報検索の結果を起点とした効率的な知識の獲得ができないという課題があった。この課題を解決するための情報検索装置を提供する。

【解決手段】 文書データベース101を検索対象として、ユーザの情報要求に基づいて情報検索手段102により基本的な情報検索を行い、取得した検索結果のリストにおける文書群を検索結果分類手段103により互いに類似した文書により構成される文書集合に分類し、特徴的単語抽出手段104と特徴的関係抽出手段105により検索結果の各文書集合から特徴的な単語および特徴的な関係を抽出し、分類されたグループと、抽出された単語及び関係に基づいてユーザに対して操作可能な出力画面の情報を出力情報編集手段106により生成する。



【特許請求の範囲】

【請求項1】 予め構築された文書データベースをユーザにより入力される情報要求に従って検索する情報検索装置であって、前記文書データベースを検索対象として、前記ユーザの情報要求に基づいて基本的な情報検索を行う情報検索手段、該情報検索手段により取得した検索結果のリストにおける文書群を文書集合に分類する検索結果分類手段、前記文書群それぞれから特徴的な単語を抽出する特徴的単語抽出手段、前記文書群それぞれから特徴的な単語の関係を抽出する特徴的関係抽出手段、前記検索結果分類手段と、前記特徴的単語抽出手段と、前記特徴的関係抽出手段とから得られた結果に基づいて、前記ユーザに対して操作可能な出力画面の情報を生成する出力情報編集手段を有することを特徴とする情報検索装置。

【請求項2】 予め構築された文書データベースをユーザにより入力される情報要求に従って検索する情報検索装置であって、前記文書データベースを検索対象として、前記ユーザの情報要求に基づいて基本的な情報検索を行う情報検索手段、該情報検索手段により取得した検索結果の文書群それぞれから特徴的な単語を抽出する特徴的単語抽出手段、該特徴的単語抽出手段により抽出した特徴的な単語に基づき、文書群を互いに類似した文書集合に分類する検索結果分類手段、前記文書群それぞれから特徴的な単語の関係を抽出する特徴的関係抽出手段、前記検索結果分類手段と、前記特徴的関係抽出手段とから得られた結果に基づいて、前記ユーザに対して操作可能な出力画面の情報を生成する出力情報編集手段を有することを特徴とする情報検索装置。

【請求項3】 予め構築された文書データベースをユーザにより入力される情報要求に従って検索する情報検索装置であって、前記文書データベースを検索対象として、前記ユーザの情報要求に基づいて基本的な情報検索を行う情報検索手段、該情報検索手段により取得した検索結果の文書群それぞれから特徴的な単語を抽出する特徴的単語抽出手段、前記文書群それぞれから特徴的な単語の関係を抽出する特徴的関係抽出手段、前記特徴的単語抽出手段および特徴的関係抽出手段により抽出した特徴的な単語および特徴的な単語の関係に基づき、文書群を互いに類似した文書集合に分類する検索結果分類手段、該検索結果分類手段から得られた結果に基づいて、前記ユーザに対して操作可能な出力画面の情報を生成する出力情報編集手段を有することを特徴とする情報検索装置。

【請求項4】 出力情報編集手段が、出力画面の表示と同時に、ユーザの次の行動を入力可能な画面情報を表示する手段を含む請求項1ないし3のいずれかに記載の情報検索装置。

【請求項5】 検索結果分類手段が、検索結果分類手段により分類された文書集合に対して、再度分類を行う手

段を含む請求項1ないし3のいずれかに記載の情報検索装置。

【請求項6】 情報検索手段が、特徴的単語抽出手段および特徴的関係抽出手段により抽出した特徴的な単語および特徴的な単語の関係の少なくとも一つを情報要求として再度検索を実行する手段を含む請求項1ないし3のいずれかに記載の情報検索装置。

【請求項7】 予め構築された文書データベースをユーザにより入力される情報要求に従って検索する情報検索プログラムを格納した記憶媒体であって、前記文書データベースを検索対象として、前記ユーザの情報要求に基づいて基本的な情報検索を行う情報検索プロセス、該情報検索プロセスにより取得した検索結果のリストにおける文書群を文書集合に分類する検索結果分類プロセス、前記文書群それぞれから特徴的な単語を抽出する特徴的単語抽出プロセス、前記文書群それぞれから特徴的な単語の関係を抽出する特徴的関係抽出プロセス、前記検索結果分類プロセスと、前記特徴的単語抽出プロセスと、前記特徴的関係抽出プロセスとから得られた結果に基づいて、前記ユーザに対して操作可能な出力画面の情報を生成する出力情報編集プロセスを有することを特徴とする記憶媒体。

【請求項8】 予め構築された文書データベースをユーザにより入力される情報要求に従って検索する情報検索プログラムを格納した記憶媒体であって、前記文書データベースを検索対象として、前記ユーザの情報要求に基づいて基本的な情報検索を行う情報検索プロセス、該情報検索プロセスにより取得した検索結果の文書群それぞれから特徴的な単語を抽出する特徴的単語抽出プロセス、該特徴的単語抽出プロセスにより抽出した特徴的な単語に基づき、文書群を互いに類似した文書集合に分類する検索結果分類プロセス、前記文書群それぞれから特徴的な単語の関係を抽出する特徴的関係抽出プロセス、前記検索結果分類プロセスと、前記特徴的関係抽出プロセスとから得られた結果に基づいて、前記ユーザに対して操作可能な出力画面の情報を生成する出力情報編集プロセスを有することを特徴とする記憶媒体。

【請求項9】 予め構築された文書データベースをユーザにより入力される情報要求に従って検索する情報検索プログラムを格納した記憶媒体であって、前記文書データベースを検索対象として、前記ユーザの情報要求に基づいて基本的な情報検索を行う情報検索プロセス、該情報検索プロセスにより取得した検索結果の文書群それぞれから特徴的な単語を抽出する特徴的単語抽出プロセス、前記文書群それぞれから特徴的な単語の関係を抽出する特徴的関係抽出プロセス、該特徴的単語抽出プロセスおよび前記特徴的関係抽出プロセスにより抽出した特徴的な単語および特徴的な単語の関係に基づき、文書群を互いに類似した文書集合に分類する検索結果分類プロセス、該検索結果分類プロセスから得られた結果に基づ

いて、前記ユーザに対して操作可能な出力画面の情報を生成する出力情報編集プロセスを有することを特徴とする記憶媒体。

【請求項10】 出力情報編集プロセスが、出力画面の表示と同時に、ユーザの次の行動を入力可能な画面情報を表示するプロセスを含む請求項7ないし9のいずれかに記載の記憶媒体。

【請求項11】 検索結果分類プロセスが、検索結果分類プロセスにより分類された文書集合に対して、再度分類を行うプロセスを含む請求項7ないし9のいずれかに記載の記憶媒体。

【請求項12】 情報検索プロセスが、特徴的な単語抽出プロセスおよび特徴的な単語抽出プロセスにより抽出した特徴的な単語および特徴的な単語の少なくとも一つを情報要求として再度検索を実行するプロセスを含む請求項7ないし9のいずれかに記載の記憶媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】この発明は、文書データベースをユーザの情報要求に従って検索する情報検索装置において、検索結果を自動分類し、検索結果の文書群において特徴的な単語及び特徴的な関係を求めることにより、ユーザの対話的な情報検索行動を支援し、効率的に知識獲得をすることができる出力情報を提供する情報検索装置及び情報検索プログラムを格納した記憶媒体に関するものである。

【0002】

【従来の技術】従来の情報検索装置においては、ユーザの情報検索要求に対して検索された文書データベース中の各文書について、ユーザの情報要求に対する適合度を計算し、この適合度によってソートされた文書のリストをユーザに提供することが通常行われている。

【0003】

【発明が解決しようとする課題】しかしながら、上記の従来の情報検索装置では、この適合度は必ずしもユーザの意向や直感を反映するものであるとは限らない。また、情報検索結果は単なる文書リストであるため、情報検索結果を起点とするユーザの効率的な知識獲得を支援するには十分ではないという問題がある。

【0004】また、近年のインターネット上のサーチエンジンに顕著のように、膨大な文書群を対象とする場合、検索によって得られる文書リストは膨大なものになる場合も多く、各文書リストの要素を全てユーザがチェックすることは、事実上不可能であるという問題がある。

【0005】この発明は、上記のような問題点を鑑みながら、情報検索の結果を起点とした効率的な知識獲得ができる、対話的な情報検索装置および情報検索プログラムを格納した記憶媒体の提供を目的とする。

【0006】

【課題を解決するための手段】この発明に係る第1の情報検索装置は、予め構築された文書データベースをユーザにより入力される情報要求に従って検索する情報検索装置であって、前記文書データベースを検索対象として、前記ユーザの情報要求に基づいて基本的な情報検索を行う情報検索手段、該情報検索手段により取得した検索結果のリストにおける文書群を文書集合に分類する検索結果分類手段、前記文書群それぞれから特徴的な単語を抽出する特徴的な単語抽出手段、前記文書群それぞれから特徴的な単語の関係を抽出する特徴的な単語抽出手段、前記検索結果分類手段と、前記特徴的な単語抽出手段と、前記特徴的な単語抽出手段とから得られた結果に基づいて、前記ユーザに対して操作可能な出力画面の情報を生成する出力情報編集手段を有するものである。

【0007】この発明に係る第2の情報検索装置は、予め構築された文書データベースをユーザにより入力される情報要求に従って検索する情報検索装置であって、前記文書データベースを検索対象として、前記ユーザの情報要求に基づいて基本的な情報検索を行う情報検索手段、該情報検索手段により取得した検索結果の文書群それぞれから特徴的な単語を抽出する特徴的な単語抽出手段、該特徴的な単語抽出手段により抽出した特徴的な単語に基づき、文書群を互いに類似した文書集合に分類する検索結果分類手段、前記文書群それぞれから特徴的な単語の関係を抽出する特徴的な単語抽出手段、前記検索結果分類手段と、前記特徴的な単語抽出手段とから得られた結果に基づいて、前記ユーザに対して操作可能な出力画面の情報を生成する出力情報編集手段を有するものである。

【0008】この発明に係る第3の情報検索装置は、予め構築された文書データベースをユーザにより入力される情報要求に従って検索する情報検索装置であって、前記文書データベースを検索対象として、前記ユーザの情報要求に基づいて基本的な情報検索を行う情報検索手段、該情報検索手段により取得した検索結果の文書群それぞれから特徴的な単語を抽出する特徴的な単語抽出手段、前記文書群それぞれから特徴的な単語の関係を抽出する特徴的な単語抽出手段、該特徴的な単語抽出手段および特徴的な単語の関係に基づき、文書群を互いに類似した文書集合に分類する検索結果分類手段、該検索結果分類手段から得られた結果に基づいて、前記ユーザに対して操作可能な出力画面の情報を生成する出力情報編集手段を有するものである。

【0009】この発明に係る第4の情報検索装置は、前記第1ないし第3の情報検索装置のいずれかにおいて、出力情報編集手段が、出力画面の表示と同時に、ユーザの次の行動を入力可能な画面情報を表示する手段を含むものである。

【0010】この発明に係る第5の情報検索装置は、前

記第1ないし第3の情報検索装置のいずれかにおいて、検索結果分類手段が、検索結果分類手段により分類された文書集合に対して、再度分類を行う手段を含むものである。

【0011】この発明に係る第6の情報検索装置は、前記第1ないし第3の情報検索装置のいずれかにおいて、情報検索手段が、特徴的単語抽出手段および特徴的関係抽出手段により抽出した特徴的な単語および特徴的な単語の関係の少なくとも一つを情報要求として再度検索を実行する手段を含むものである。

【0012】この発明に係る第1の記憶媒体は、予め構築された文書データベースをユーザにより入力される情報要求に従って検索する情報検索プログラムを格納した記憶媒体であって、前記文書データベースを検索対象として、前記ユーザの情報要求に基づいて基本的な情報検索を行う情報検索プロセス、該情報検索プロセスにより取得した検索結果のリストにおける文書群を文書集合に分類する検索結果分類プロセス、前記文書群それぞれから特徴的な単語を抽出する特徴的単語抽出プロセス、前記文書群それぞれから特徴的な単語の関係を抽出する特徴的関係抽出プロセス、前記検索結果分類プロセスと、前記特徴的単語抽出プロセスと、前記特徴的関係抽出プロセスとから得られた結果に基づいて、前記ユーザに対して操作可能な出力画面の情報生成する出力情報編集プロセスを有するものである。

【0013】この発明に係る第2の記憶媒体は、予め構築された文書データベースをユーザにより入力される情報要求に従って検索する情報検索プログラムを格納した記憶媒体であって、前記文書データベースを検索対象として、前記ユーザの情報要求に基づいて基本的な情報検索を行う情報検索プロセス、該情報検索プロセスにより取得した検索結果の文書群それぞれから特徴的な単語を抽出する特徴的単語抽出プロセス、該特徴的単語抽出プロセスにより抽出した特徴的な単語に基づき、文書群を互いに類似した文書集合に分類する検索結果分類プロセス、前記文書群それぞれから特徴的な単語の関係を抽出する特徴的関係抽出プロセス、前記検索結果分類プロセスと、前記特徴的関係抽出プロセスとから得られた結果に基づいて、前記ユーザに対して操作可能な出力画面の情報生成する出力情報編集プロセスを有するものである。

【0014】この発明に係る第3の記憶媒体は、予め構築された文書データベースをユーザにより入力される情報要求に従って検索する情報検索プログラムを格納した記憶媒体であって、前記文書データベースを検索対象として、前記ユーザの情報要求に基づいて基本的な情報検索を行う情報検索プロセス、該情報検索プロセスにより取得した検索結果の文書群それぞれから特徴的な単語を抽出する特徴的単語抽出プロセス、前記文書群それぞれから特徴的な単語の関係を抽出する特徴的関係抽出プロ

セス、該特徴的単語抽出プロセスおよび前記特徴的関係抽出プロセスにより抽出した特徴的な単語および特徴的な単語の関係に基づき、文書群を互いに類似した文書集合に分類する検索結果分類プロセス、該検索結果分類プロセスから得られた結果に基づいて、前記ユーザに対して操作可能な出力画面の情報生成する出力情報編集プロセスを有するものである。

【0015】この発明に係る第4の記憶媒体は、前記第1ないし第3の記憶媒体のいずれかにおいて、出力情報編集プロセスが、出力画面の表示と同時に、ユーザの次の行動を入力可能な画面情報を表示するプロセスを含むものである。

【0016】この発明に係る第5の記憶媒体は、前記第1ないし第3の記憶媒体のいずれかにおいて、検索結果分類プロセスが、検索結果分類プロセスにより分類された文書集合に対して、再度分類を行うプロセスを含むものである。

【0017】この発明に係る第6の記憶媒体は、前記第1ないし第3の記憶媒体のいずれかにおいて、情報検索プロセスが、特徴的単語抽出プロセスおよび特徴的関係抽出プロセスにより抽出した特徴的な単語および特徴的な単語の関係の少なくとも一つを情報要求として再度検索を実行するプロセスを含むものである。

【0018】

【発明の実施の形態】実施の形態1. 図1は、本発明の原理構成図である。101は予め構築された文書データベース、102は文書データベース101を検索対象として、ユーザの情報要求に基づいて基本的な情報検索を行う情報検索手段、103は情報検索手段102により取得した検索結果のリストにおける文書群を文書集合に分類する検索結果分類手段、104は検索結果の各文書から特徴的な単語のリストを抽出する特徴的単語抽出手段、105は検索結果の各文書から特徴的な単語間の関係のリストを抽出する特徴的関係抽出手段、106は検索結果分類手段103と特徴的単語抽出手段104と特徴的関係抽出手段105の結果に基づいて、ユーザに対して操作可能な出力画面の情報生成する出力情報編集手段である。

【0019】図1の原理構成において、まず、ユーザは情報要求のキーワードを情報検索手段102に入力し、文書データベース101の基本的な検索を行い、文書群のリストを得る。

【0020】次に、検索結果分類手段103においては、例えば、検索結果のリストにおける文書群を、その文書が入力された日時、文書の提供者、文書提供者の業界等、文書に入力されているキーワードに基づいて分類を実行し、分類された文書リストを得る。

【0021】また、検索結果のリストを特徴的単語抽出手段104と特徴的関係抽出手段105に転送し、検索された文書それぞれの中に含まれる特徴的な単語と特徴

10

20

30

40

50

的な単語の関係を抽出する。ここで、特徴的単語および特徴的な単語の関係とは、文書に含まれる単語および単語の関係のリストである。各単語には、その重要性を表す重み、例えば、実数値が付与されているものとする。すなわち、文書それぞれの特徴は実数値を値とし、各要素はある単語に対応したベクトルとして表現される。また、特徴的関係とは、文書に含まれる一つ以上の単語とそれらを結びリンクから構成される、データ構造を表す。単語間を結びリンクは、その両端の単語間が該当文書において密接な関係にあることを表す。より具体的には、該当文書中でそれらに文法的な係り受け関係があることや、表層的な出現位置が特定の閾値以下であること、また、文章の論理構造的に出現位置が閾値以下であること（例えば、同一文章中に出現する）などが挙げられる。これらの特徴的関係を用いて、文書それぞれの特徴は、例えば、実数値を値とし、各要素はある単語の関係に対応したベクトルとして表現される。

【0022】単語の重みとしては、情報検索の分野において従来より検討がなされており、単純な頻度、正規化された頻度や特徴的であるか否かを表す値（ $TF * IDF$ ）などが考えられる。本発明では、単語の重みとして何を使用するかについては規定しない。また、特徴的単語抽出手段104では、指示された文書の特徴をその都度計算するのではなく、文書の特徴を内部データベースにキャッシュしておくなどの効率化手段が考えられるが、本発明ではその詳細は規定しない。特徴的関係に関しても特徴的単語と同様の重みづけが考えられるが、本発明では関係の重みとして何を使用するかについては規定しない。

【0023】次に、出力情報編集手段106は、検索結果分類手段103と特徴的単語抽出手段104と特徴的関係抽出手段105の結果に基づいて、ユーザに対して操作可能な出力画面の情報を生成する。この出力画面情報は、同時に、ユーザが次の行動（再検索あるいは検索結果の再分類）を入力する画面でもある。

【0024】ユーザは、出力画面に情報要求のキーワードとして特徴的単語あるいは特徴的関係を追加して上記のプロセスを経て再検索を行うことができる。

【0025】また、抽出された特徴的単語あるいは特徴的関係あるいはこれら両方の組み合わせを用いて、検索された文書群を再分類することができる。この再分類の場合は、検索結果得られた文書リストあるいは分類された文書集合を検索結果分類手段103へ入力し、特徴的単語あるいは特徴的関係あるいはこれら両方の組み合わせを用いて、再分類し、その処理結果を出力情報編集手段106へ転送し、出力情報編集手段106は、ユーザが操作可能な上記と同様の出力画面の情報を生成する。

【0026】この実施の形態1によれば、基本的な検索結果について、特徴的単語とこの単語の特徴的関係を抽出することによって、質の高い検索結果（検索情報）が

得られ、下記（1）、（2）および（3）のようなことが対話的で高効率に実施可能になり、ユーザの知識獲得を高度に支援することが可能になる。

（1）情報検索結果が単なるリストではなく、検索結果の文書が情報要求を高度に満たす文書集合へと自動的に分類されるため、実際にアクセスする文書を決定する際の支援となる。

（2）さらに、上記で生成されたグループを1つ以上選択した結果である文書集合に対して、再度の自動分類を指示することにより、検索結果を絞り込んでいくことが可能になる。

（3）抽出された特徴的な単語及び特徴的な関係のリストを利用して、この中からユーザが適当な単語または関係またはこれらの組み合わせによって、絞り込み的な検索や、関連するトピックに関する検索（連想型検索）が可能となる。

【0027】実施の形態2. 図2は、本発明の情報検索装置における実施の形態2を示す構成図である。同図に示す構成（システム）は、入力部201、情報検索サービス部202、情報検索実行部203、文書データベース204、文書分類サービス部205、検索結果分類実行部206、特徴的単語抽出部207、特徴的関係抽出部208、出力情報編集部209、出力部210を有する。

【0028】入力部201は、ユーザ端末から送信されるシステムへの指示を受け付ける。ユーザからの指示は、検索のための情報要求、または、検索結果の再分類の指示のどちらかである。入力部201は、これらの判断を行い、前者の場合には、情報検索サービス部202へ入力された情報要求を転送し、後者の場合には、文書分類サービス部205へ入力された検索結果の再分類の指示を転送する。以下では、まず、検索のための情報要求について説明する。

【0029】情報検索サービス部202は、まず、入力部201から転送されてきた情報要求を情報検索実行部203へ転送する。情報検索実行部203は、予め構築された文書データベース204を検索対象とした情報検索を実行し、その結果として、ユーザの情報要求に対する適合度順にソートされた、文書データベース204の文書リストを情報検索サービス部202へ返却する。なお、情報検索実行部203に相当するシステムは、公知の技術により十分実現可能であるため、その詳細は問わず、入力として単語の論理結合（AND、OR、NOT）を許すことと、適合度順にソートされた文書リストを結果とすることのみ条件とする。

【0030】検索結果分類実行部206は、情報検索サービス部202より転送されてくる文書群（実際には文書データベース204における文書IDの集合）を入力として受ける。検索結果分類実行部206は、まず、特徴的単語抽出部207を呼び出すことにより、文書デー

データベース204における指定された文書における特徴的単語を得る。ここで、特徴的単語とは、文書に含まれる単語のリストであり、各単語には、その重要性を表す重み、例えば、実数値が付与されているものとする。すなわち、文書の特徴は実数値を値とし、各要素はある単語に対応しているベクトルとして表現される。

【0031】単語の重みとしては、情報検索の分野において従来より検討がなされており、単純な頻度、正規化された頻度や特徴的であるか否かを表す値(TF*IDF)などが考えられる。本発明では、単語の重みとして何を使用するかについては規定しない。また、特徴的単語抽出部207では、指示された文書の特徴をその都度計算するのではなく、文書の特徴を内部データベースにキャッシュしておくなどの効率化手段が考えられるが、その詳細は規定しない。

【0032】また、検索結果分類実行部206は、特徴的関係抽出部208を呼び出すことにより、文書データベース204における指定された文書における特徴的関係を得る。ここで、特徴的関係とは、文書に含まれる一つ以上の単語とそれらを結ぶリンクから構成される、データ構造を表す。単語間を結ぶリンクは、その両端の単語間が該当文書において密接な関係にあることを表す。より具体的には、該当文書中でそれらに文法的な係り受け関係があることや、表層的な出現位置が特定の閾値以下であること、また、文章の論理構造的に出現位置が閾値以下であること(例えば、同一文章中に出現する)などが挙げられる。これらの特徴的関係を用いて、文書の特徴は実数値を値とし、各要素はある関係に対応しているベクトルとして表現される。特徴的関係に関しても特徴的単語と同様の重みづけが考えられるが、本発明では関係の重みとして何を使用するかについては規定しない。また、特徴的関係抽出部208においても、特徴的単語抽出部207と同様、内部データベースにキャッシュしておくなどの効率化手段が考えられるが、その詳細は規定しない。

【0033】次に、検索結果分類実行部206は、入力された文書群の文書それぞれに対して求められた文書特徴ベクトルを総合することにより、図3に示すような行列を求める。当該行列の各行は文書に、各列は単語または関係に相当する。このような行列を以下では特徴行列と呼ぶ。特徴行列の構成時に、特徴的関係の重みのバランスを取るために、定数を二種類(α , β)用意し、特徴的単語を表す重みには α を、特徴的関係を表す重みには β を乗じてよい。 $(\alpha, \beta) = (1, 0)$ のときは特徴的単語のみを考慮した分類となり、 $(\alpha, \beta) = (0, 1)$ のときは特徴的関係のみを考慮した分類となる。

【0034】つぎに、検索結果分類実行部206は、特徴行列に基づき互いに類似した文書を文書集合に分類する。特徴行列の近いもの同士をグループ化することによ

り、文書の自動分類を行う方法として、クラスタリングと呼ばれる手法が知られており、いくつかのアルゴリズムが提案されている(参考文献例:E. Rasmussen: Clustering Algorithms, in W. B. Frakes, R. Baeza-Yates, editors, Information Retrieval, Prentice Hall, 1992)。この発明における検索結果分類実行部206の採用するクラスタリングのアルゴリズムについては規定しないが、これらの処理の結果、特徴行列が互いに類似した文書をグループ化することが可能となる。特徴行列は特徴的単語と特徴的関係から構成されるため、結果として内容的に互いに類似した文書をグループ化することができる。

【0035】次に、検索結果分類実行部206は、分類された結果に応じて、図4に示すような、各グループにおいて特徴的な単語または関係のリストも求めるものとする。その方法の一つとしては、多くのクラスタリングアルゴリズムではグループ毎にクラスタ中心という仮想的な特徴群に基づいて分類処理を行うため、各グループにおけるクラスタ中心から重みの大きい特徴を取り出して用いることが考えられる。また、別の方法として、各クラスタに分類された文書から再度特徴的単語や特徴的関係を抽出し、重みの大きい特徴を取り出して用いることも考えられる。

【0036】特徴の取り出し方としては、リストの大きさ(単語または関係の数)を陽に指定して大きいものから順に取得する方法や、ある一定の値以上の重みを持つ単語または半径のみを対象としてリストを構成する方法が考えられる。また、特徴行列の構成と同様、定数を二種類用意し、それぞれ特徴的単語を表す重みと特徴的関係を表す重みに乗じた後、リストを構成してもよい。これらの定数は特徴行列の生成時に用いた定数とは独立に決定してもよい。

【0037】出力情報編集部209は、検索結果分類実行部206から下記(1)及び(2)のデータを受け取り、ユーザによるインタラクティブな情報検索行動を支援するための出力画面情報を生成する。この出力画面情報は、同時に、ユーザが次の再分類または再検索の行動を入力する画面でもある。

(1) グループに属する文書のリスト

(2) グループを特徴付ける単語及び関係のリストと各々の重み

【0038】出力部210は、出力情報編集部209から転送されてきた画面情報をユーザ端末へと転送する。

【0039】ユーザは、出力画面に情報要求のキーワードとして特徴的単語あるいは特徴的関係を追加し、上記のプロセスを経て再検索を行うことができる。

【0040】また、ユーザは検索された文書集合を再分類するために、再分類の指示を入力部201に指示し、

10

20

30

40

50

入力部201は再分類の指示を文書分類サービス部205へ転送する。文書分類サービス部205は、再分類の指示に基づいて、再分類の対象となる文書集合を出力情報編集部209に求め、この文書集合を検索結果分類実行部206へ転送し、さらに、その処理結果を出力情報編集部209に転送し、ユーザが操作可能な上記と同様の出力画面の情報を生成する。

【0041】この実施の形態2においては、検索された文書リスト中文書群の自動分類に先だって特徴的単語および特徴的関係を抽出し、この抽出された特徴的単語および特徴的関係を利用して検索された文書群を類似した文書集合に分類しているので、極めて質の高い検索結果（検索情報）が得られ、対話的で高効率に、実際にアクセスする文書をユーザが決定するのを支援することができる。また、再度の自動分類を指示することによる検索結果の絞り込みができる。また、自動分類に至る過程で抽出される特徴的な単語及び特徴的な関係のリストを利用した絞り込み的な検索や、関連するトピックに関する検索（連想型検索）が可能となる。

【0042】なお、検索された文書リスト中文書群の自動分類に先だって特徴的単語および特徴的関係を抽出し、この抽出された特徴的単語を利用して検索された文書群を類似した文書集合に分類し、分類された文書集合と別に抽出された特徴的関係を研修して出力することにより、出力までの時間を短縮することができる。

【0043】

【実施例】以下、図面に基づき、具体的な実施例により本発明を説明するが、本発明が上この実施例に限定されるものではなく、特許請求の範囲内で種々変更・応用が可能である。

【0044】この実施例は、本発明の情報検索装置をWWW(World Wide Web)上の検索エンジンに適用した場合である。

【0045】図5は、本発明の一実施例における検索要求入力画面の例を示しており、ユーザ端末に表示される初期画面の例である。この画面例は、まず、情報要求を表す語を入力する領域が最上部に設定されており、ユーザは、「インターネット」、「電子商取引」の2語からなる情報要求を入力したことを示している。その次の行では、検索方式に関する基本的な設定が行えるようになっており、複数入力された検索語をAND条件で結んで検索することが指定されている。その次の行では、検索結果のリスト表示に関する設定が行えるようになっており、検索語を多く含む文書から10件ずつのリストを表示するように指定している。その次の行では、検索結果のクラスタ処理に関する設定が行えるようになっており、クラスタ数自動でクラスタ処理を行うように指定している。以下では、この入力例に即し、図2を参照して説明を行う。

【0046】入力部201は、ユーザ端末から送信されてくる上記のような入力画面の要求のタイプに従って、ユーザの要求を情報検索サービス部202、または、文書分類サービス部205へ転送する。この実施例の場合は、情報検索のための情報要求であるので、情報検索サービス部202へ要求を転送する。情報検索サービス部202は、転送されてきた情報要求と検索条件から検索式を生成する。この実施例の情報要求及び検索条件からは、（インターネット）AND（電子商取引）なる検索式が生成され、この検索式は情報検索実行部203へ転送される。

【0047】図6は、この実施例における情報検索結果の文書リストの例を示し、上記の検索式によって情報検索実行部203が文書データベース204を検索対象として検索を行った結果である。情報検索実行部203に適用されるような通常のテキスト検索システムは、ここに示されたような情報以外の情報を返却することも可能であるが、図7では、以下の説明に必要な最小限の情報のみを示す。すなわち、情報検索結果の各文書に対しては、その文書データベース204内における文書ID（この例では3桁の数）、文書のタイトル（説明を分かり易くするために導入した）が返却されたものとしている。

【0048】情報検索サービス部202は、図6に示された検索結果を検索結果分類実行部206へ転送する。検索結果分類実行部206は、まず、特徴的単語抽出部207を呼び出すことにより、転送されてきた文書の特徴的な単語を得る。また、検索結果分類実行部206は、特徴的関係抽出部208を呼び出すことにより、転送されてきた文書の特徴的な関係を得る。図7は、得られた文書特徴ベクトルの例であり、図7の第1要素（ID=001、タイトル=電子商取引の持つ危険性）について得られた例である。説明を簡単にするため、この文書の2つの特徴的単語である（詐欺）：3、（改ざん）：2及び1つの特徴的関係である（クレジット番号→不正使用）：2のみを示す。ここで、実数は強さを表している。また、ここに挙げた特徴的関係は係り受け関係に基づくものであり、→は、係る単語→係られる単語の関係を示している。

【0049】次に、検索結果分類実行部206が、入力された文書集合の各要素である文書に対して求められた文書特徴ベクトルを総合することにより、図3に示すような行列を求める。図8は、この実施例における特徴行列の例であり、検索結果リストに対する特徴行列を説明するための図である。図8は、前述の図6の検索結果リストに対する特徴行列を示しており、同図では、図7と同様に説明を簡単にするため、検索結果で得られた10の文書に対し主要な特徴的単語5個と特徴的関係2個のみを示したが、実際においては、特徴的単語と特徴的関係の数はこれにとどまるものではない。

【0050】次に、検索結果分類実行部206は、図8の特徴行列に対してクラスタリングアルゴリズムを実行する。前述したように、いくつかのクラスタリングアルゴリズムが提案されているので、本実施例における検索結果分類実行部206は、適当なアルゴリズムを実装していると仮定する。図9は、図8の特徴行列に対してクラスタリングアルゴリズムを実行した結果の分類結果行列を説明するものである。図9に示したように、この例においては、10の文書が2つのグループ（1つは6つの文書からなり、もう1つは4つの文書からなる）へ自動分類されている。

【0051】図10は、検索結果分類実行部206の処理結果として、分類結果行列と同時に得られる特徴的単語および特徴的関係のリストを説明するための図である。図10において、第1の文書グループにおける特徴的単語および特徴的な関係がその重みと共に示されている。

【0052】上記のような検索結果分類実行部206の処理結果は、出力情報編集部209へ転送される。出力情報編集部209は、転送されてきたデータに基づいて、ユーザによる対話的な情報検索行動を支援するための出力画面情報を生成する。この出力画面は、同時に、ユーザが次に行う再検索あるいは再分類の行動を入力するための画面でもある。出力部210は、出力情報編集部209から転送されてきた出力画面情報をユーザの端末へ転送する。

【0053】図11及び図12は、この実施例における出力画面情報の例を示している。これらの図11および図12は、出力部210によりユーザの端末に転送される具体的な出力画面情報の例である。

【0054】図11は、上記の検索結果として生成された再検索のための入力フォームと、検索された文書のリストを表す。図11において、上部に再検索のための入力フォームが、その下に検索された文書のタイトルが列挙されている。各タイトル部をクリックすることで具体的な文書の内容を参照することができる。図11の入力フォームにおいて、検索方式、結果表示、クラスタ処理等の選択メニューは、図5に示したものと同様のものであり、検索語のフィールドには先に入力した語が予め入力されているが、ユーザはこれらの語を編集もしくは語を追加して再検索を行うことが容易となっている。また、自動分類処理の段階で抽出された特徴的な単語および特徴的な関係のうち主たるものが表示されている。これらの語に付与されたボタンをクリックすることで、検索語のフィールドに当該語を入力し、再検索を容易に行うことが可能になる。

【0055】図12は、自動分類の結果として生成された文書グループ（画面ではクラスタと表記）の情報を表示している。同図において、4文書からなるクラスタ1と、6文書からなるクラスタ2が生成されたことが示さ

れている。各クラスタ番号の次に表示されているのは、当該クラスタにおける特徴的な単語および特徴的な関係、及びそれらの当該クラスタにおける強さである。その次に列挙されているのは、当該クラスタに分類された記事のタイトルを表す。タイトル部をクリックすることで具体的な文書の内容を参照することができる。各クラスタの先頭のボタンを選択し、末尾の再クラスタリングボタンをクリックすることにより、当該クラスタに含まれる文書を再クラスタリングし、より詳細な情報を得ることができる。

【0056】

【発明の効果】上述のように、本発明によれば、情報検索結果の自動分類や、検索結果の部分集合に対する再自動分類による検索結果の構造化、自動分類の過程で抽出された特徴的な単語および関係を組み合わせることによる次の段階の検索支援が可能となり、これらは、情報検索に基づくユーザの知識獲得を支援する。

【図面の簡単な説明】

【図1】 本発明における情報探索装置の実施の形態1を説明する構成図である。

【図2】 本発明における情報探索装置の実施の形態2を説明する構成図である。

【図3】 本発明の実施の形態2における、文書リストの特徴行列の例を示す図である。

【図4】 本発明の実施の形態2における、文書グループの特徴的な単語または特徴的な関係のリストの例を示す図である。

【図5】 本発明の一実施例における、検索要求入力画面の例を示す図である。

【図6】 本発明の一実施例における、情報検索結果の文書リストの例を示す図である。

【図7】 本発明の一実施例における、文書特徴ベクトルの例を示す図である。

【図8】 本発明の一実施例における、文書リストの特徴行列の例を示す図である。

【図9】 本発明の一実施例における、分類結果行列の例を示す図である。

【図10】 本発明の一実施例における、特徴的単語および特徴的関係の例を示す図である。

【図11】 本発明の一実施例における、出力画面の一例を示す図である。

【図12】 本発明の一実施例における、出力画面の他の例を示す図である。

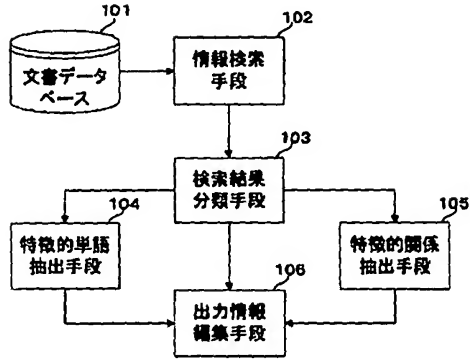
【符号の説明】

101、204 文書データベース、102 情報検索手段、103 検索結果分類手段、104 特徴的単語抽出手段、105 特徴的関係抽出手段、106 出力情報編集手段、201 入力部、202 情報検索サービス部、203 情報検索実行部、205 文書分類サービス部、206 検索結果分類実行部、207 特徴

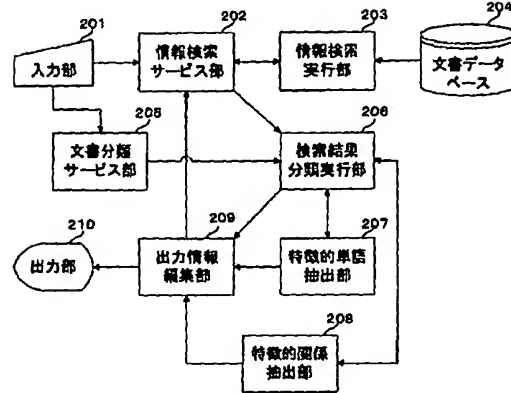
的単語抽出部、208 特徴的關係抽出部、209 出

力情報編集部、210 出力部。

【図1】



【図2】



【図3】

	T1	T2	T3	...	Tm
D1	W11	W12	W13	...	W1m
D2	W21	W22	W23	...	W2m
D3	W31	W32	W33	...	W3m
...
Dn	Wn1	Wn2	Wn3	...	Wnm

D_i : 文書 i , T_j : 単語または関係 j , W_{ij} : 文書 i における単語または関係 j の重み

【図4】

T1:W11, T3:W13, T4:W14, T7:W17, ...

文書グループから特徴的な単語または関係としてT1, T3, T4, T7, ...を抽出した場合

【図5】

検索語: →

検索方式:

結果表示:

クラス処理: ☒ クラス数:

【図6】

ID	タイトル
001	電子商取引の持つ危険性
002	ビジネスとインターネット
003	電子マネー
004	電子商取引の法務と税務
005	インターネットトレンド
006	有カベンチャーが電子商取引に出資
007	電子商取引の関税ゼロ、日米首脳が声明
008	電子商取引とは
009	電子商取引とセキュリティ
010	通産省が電子商取引での消費者保護を検討

【図7】

(詐欺):3, (改竄):2, (クレジット番号→不正使用):2

【図8】

	ID	T1 詐欺	T2 改竄	T3 オンライン ショッピング	T4 暗号化	T5 電子マネー	T6 クレジット番号 →不正使用	T7 EC →入門
D1	001	3	2	0	0	0	4	0
D2	002	0	0	5	1	2	0	1
D3	003	0	0	2	2	6	0	2
D4	004	2	1	0	1	1	2	0
D5	005	0	0	2	1	1	0	0
D6	006	0	0	3	1	2	0	0
D7	007	0	0	1	1	1	0	0
D8	008	0	0	2	2	5	0	3
D9	009	1	1	0	5	3	1	0
D10	010	3	1	2	2	2	3	0

【図9】

	文書	T1 詐欺	T2 改竄	T3 オンライン ショッピング	T4 暗号化	T5 電子マネー	T6 クレジット番号 →不正使用	T7 EC →入門
G1	D1,D4,D9, D10	9	5	2	8	6	10	0
G2	D2,D3,D5, D6,D7,D8	0	0	15	8	17	0	6

【図10】

(詐欺):9, (改竄):5, (オンラインショッピング):2, (暗号化):8, (電子マネー):6, (クレジット
番号→不正使用):10, (EC→入門):0

【図11】

検索語:

検索方式:

結果表示:

クラスタ処理: クラスタ数:

☐ 詐欺 ☐ 改竄 ☐ オンラインショッピング ☐ 暗号化 ☐ 電子マネー
☐ クレジット番号→不正使用 ☐ EC→入門

10件の文書が見つかりました。

1. 「電子商取引の持つ危険性」
2. 「ビジネスとインターネット」
3. 「電子マネー」
4. 「電子商取引の法務と税務」
5. 「インターネットトレンド」
6. 「有力ベンチャーが電子商取引に出資」
7. 「電子商取引の関税ゼロ、日米首脳が声明」
8. 「電子商取引とは」
9. 「電子商取引とセキュリティ」
10. 「通産省が電子商取引で消費者保護を検討」

【図12】

2個のクラスタを生成しました。

□ クラスタ1 (クレジット番号→不正使用:10 詐欺:9 暗号化:8 電子マネー:6 改竄:5 オンラインショッピング:2)

- 「電子商取引の持つ危険性」
- 「電子商取引の法務と税務」
- 「電子商取引とセキュリティ」
- 「通産省が電子商取引で消費者保護を検討」

□ クラスタ2 (電子マネー:17 オンラインショッピング:15 暗号化:8 EC→入門:8)

- 「ビジネスとインターネット」
- 「電子マネー」
- 「インターネットトレンド」
- 「有力ベンチャーが電子商取引に出資」
- 「電子商取引の関税ゼロ、日米首脳が声明」
- 「電子商取引とは」

フロントページの続き

(72)発明者 小中 裕喜

東京都千代田区丸の内二丁目2番3号 三
菱電機株式会社内

Fターム(参考) 5B075 ND03 NG31 NR12 PQ02